

Second-order Temporal Pooling for Action Recognition

Anoop Cherian · Stephen Gould

Abstract Most successful deep learning models for action recognition generate predictions for short video clips, which are later aggregated into a longer time-frame action descriptor by computing a statistic over these predictions. Zeroth (max) or first order (average) statistic are commonly used. In this paper, we explore the benefits of using second-order statistics. Specifically, we propose a novel end-to-end learnable action pooling scheme *temporal correlation pooling* that generates an action descriptor for a video sequence by capturing the similarities between the temporal evolution of per-frame CNN features across the video. Such a descriptor, while being computationally cheap, also naturally encodes the co-activations of multiple CNN features, thereby providing a richer characterization of actions than their first-order counterparts. We also propose higher-order extensions of this scheme by computing correlations after embedding the CNN features in a reproducing kernel Hilbert space. We provide experiments on four standard and fine-grained action recognition datasets. Our results clearly demonstrate the advantages of higher-order pooling schemes, achieving state-of-the-art performance.

1 Introduction

The recent resurgence of efficient deep learning architectures has facilitated significant advances in several fundamental problems in computer vision, including human action recognition [60, 19, 68, 35, 16]. However, despite these breakthroughs, the problem of action recognition is far from solved and continues to be challenging in a general setting. Real-world actions are often different from each other in very subtle ways (e.g., *washing plates* versus *washing hands*), may have strong appearance variations (e.g., *slicing cucum-*

bers versus *slicing tomatoes*), may involve significant occlusions of objects or human-body parts, may involve background activities, may use hard-to-detect objects (such as knives, peelers, etc.), and may happen over varying durations or at different rates. In this paper, we explore various schemes to address some of these issues. While our schemes are applicable in a general setting, we also explore their suitability in a *fine-grained setting* that is comprised of activities having low inter-class diversity, and high intra-class diversity [56].

Unsurprisingly, the recent approaches for activity recognition are based on deep learning algorithms [60, 19, 75, 63, 83]. Most of the successful algorithms for this problem are extensions of convolutional neural network (CNN) models originally designed for image-based recognition tasks [38]. However, in contrast to images, video data is volumetric, and thus extending such image-based models leads to huge computational and memory overheads, which are difficult to be addressed under currently available hardware platforms. A workaround, that is often found to be promising, is to reduce the video-based recognition problem into simpler image-sized subproblems, the results from these sub-problems are later collated in a fusion layer to generate predictions for the full video. While single frames might be insufficient to capture the actions effectively as they lack any temporal aspect, using longer clips demands more CNN parameters, and thus requires more training data and computational resources. As a result, popular deep action classifiers are trained on tiny sub-sequences (of 10–16 frames); the predictions from which are later pooled to generate sequence level predictions [60].

Typically, max-pooling or average pooling of the sub-sequence level predictions is used [60, 35, 74]. Although, such pooling operations are easy to implement and fast to compute, they ignore valuable higher-level information contained in the independent predictions that could improve the recognition [12, 37, 51, 13]. For example, in the context

of fine-grained recognition, let us consider two activities: *washing plates* and *wiping plates*. As is clear, discriminating the two actions is not easy due to their appearance similarities. Suppose sequences for the former also incorporate another sub-activity, say *running water from tap*. While, there could be a few frames from this activity (say at the beginning or end of sequences) for *wiping plates*, a first-order pooling scheme may detect this activity along with detections for the confusing *wiping plates* and *washing plates*. However, if we compute the co-occurrences of classifier scores, it is clear that both *washing plates* and *running water from tap* demonstrate a high temporal correlation. Given that this correlation is absent for *wiping plates*, it is easy for an action classifier trained on such correlations to discriminate the two actions. While, this is a didactic example, the main idea is that such co-occurrence relations provide valuable cues for effective recognition. Towards this end, the goal of this paper is to explore second-order co-occurrences of classifier predictions for improving action recognition.

In this paper, we propose *temporal correlation pooling* (TCP), a second-order feature pooling scheme, that takes as input a temporal sequence of CNN features (from any intermediate layer), one per video frame (Section 3.4). Each dimension of the features across time can be viewed as a *feature trajectory* corresponding to the temporal evolution of activations of the respective CNN filters. TCP summarizes these trajectories into a symmetric positive definite (SPD) matrix, each entry of this matrix capturing the similarities between such trajectories. There are several benefits that such a representation offers in contrast to prior approaches, namely (i) SPD matrices, although spanning a Euclidean subspace, are often viewed through the lens of Riemannian geometry, which offers rich non-linear distance measures for similarity computations that may help extract useful cues for recognition, (ii) SPD matrices can be naturally viewed as Mercer kernels, and similarities could be computed after embedding the feature trajectories in an infinite dimensional reproducing kernel Hilbert space (RKHS), thereby enhancing their representational power, and (iii) incorporating prior information is straightforward via sum or product kernels to the SPD kernel.

On the downside, TCP descriptors are quadratic in the size of the input features, which may be infeasible when high-dimensional features from intermediate CNN layers are used. To circumvent this issue, we propose block-diagonal correlation matrix approximations using product quantization and model averaging. Each block matrix in the resulting representation is a small positive definite matrix and thus the above recognition framework can be directly applied.

Another shortcoming of our pooling scheme is related to the confidence of the underlying CNN model; if this model is not effective in providing reliable features, the generated descriptor will be ineffective for recognition. Although, we

base our CNN on the popular two-stream model (using RGB frames for context and short stack of optical flow images for representing action dynamics), such a model lacks in two aspects: (i) long-range temporal evolution of actions, and (ii) coupling between appearance and dynamics. While, there are several recent methods that try to address these weaknesses [83, 75, 19], we propose a simpler workaround that is computationally very cheap, while empirically beneficial. Specifically, we propose a novel video representation dubbed *Stacked Mean of Absolute Image Differences* (SMAID) that is based on averaging and stacking the absolute differences of a small set of consecutive video frames. In contrast to CNN classifiers trained on single frames or flow images, our experiments show that SMAID trained classifiers demonstrate superior frame-level action predictions, especially in the context of fine-grained recognition. Incorporating this representation, we propose a three-stream end-to-end learnable CNN framework consisting of a single frame RGB stream for action context, ten-channel optical flow stream for capturing local dynamics, and a SMAID stream computed on up to 60 frames, capturing long-range dynamics (Section 4).

We provide experiments (Section 7) on four widely-used action recognition datasets to substantiate the effectiveness of our proposed schemes. Our results demonstrate that while the SMAID image representation and the correlation pooling schemes demonstrate significant gains on the fine-grained task (about 4–6%) as we expected. Surprisingly, they also lead to state-of-the-art results on general action recognition datasets.

Before moving on, we summarize the main contributions of this paper.

- We propose a novel second-order pooling scheme, dubbed temporal correlation pooling (TCP)
- We propose a kernelized variant of this pooling scheme by embedding the CNN features in an RKHS, dubbed kernelized correlation pooling (KCP)
- We address the scalability of TCP when using larger CNN features via our block-diagonal kernelized correlation pooling (BKCP).
- To boost frame-level CNN predictions we propose an enhanced frame-level video representation called SMAID.
- We propose a novel three-stream CNN action recognition model, that learns actions fusing appearance (single RGB frames), short-term (stack of optical flow), and long-term (SMAID) cues.
- We present an end-to-end learnable variant of our CNN by providing expressions for back-propagating the gradients of a classification loss computed using TCP descriptors.
- We provide extensive experimental comparisons on four benchmark datasets demonstrating state-of-the-art performance.

2 Related Work

There is an enormous breadth of approaches aimed at tackling the problem of activity recognition. We restrict attention in this literature review to methods that have similarities to ours and refer the interested reader to recent surveys [26, 8] for a detailed study of this topic.

Hand-crafted Features: Typically, in this class of methods, features derived from spatio-temporal interest points, such as dense trajectories, HOG, SIFT, HOF, etc., are extracted from regions of interest and combined to train a discriminative classifier for action recognition. Popular methods, such as those of Wang et al. [72] and Laptev [42], belong to this category. There have been extensions of these methods to use second-order statistical information of features via resorting to Fisher vectors (FV) in [71, 59, 48] and stacks of FVs [50]. While we also employ higher-order statistics, we differ from these techniques in the way we encode this information. Specifically, FVs are the parameter gradients of data modeled using a Gaussian mixture models (GMM). In contrast, our method assumes the underlying CNN implicitly captures the distribution of feature vectors, and uses the empirical covariance matrix of the probabilistic evolution of classifier scores as a representation for data. Our experiments demonstrate that the proposed representation captures complementary cues to FVs, and the synergy that comes from combining our TCP encoding with FVs results in improved accuracy (Section 7).

Deep Learning Methods: It is by now well-known that learning features in a data-driven way using deep learning can lead to better action representation [38, 60, 34, 68, 16, 83]. However, as alluded to above, scarcity of annotated video data, concomitant to the demand for expensive computational resources, makes adaptability of existing machine learning algorithms to this data modality challenging; thereby demanding efficient video representations. One of the most successful of deep learning methods for action recognition is the two-stream CNN model proposed in [60], which decouples the spatial and temporal streams, thereby learning context and action dynamics separately. These streams are trained densely and independently; and at test time, their predictions are pooled. There have been extensions to this basic architecture using deeper networks and fusion of intermediate CNN layers [19, 18, 18, 75, 74]. We also follow this trend and use a two-stream model as our baseline framework. However, we differ from these techniques in the way we use the CNN features for action recognition (first-order versus second-order). In addition, we also propose a novel three-stream CNN architecture using our SMAID image representation.

We also note that there have been several other deep learning models devised for action modeling such as using 3D convolutional filters [68], recurrent neural networks [2],

long-short term memory networks [16, 83], and large scale video classification architectures [35]. These models demand huge collections of videos for effective training, which may not be available (e.g., for fine-grained activity tasks). Further, training such models is also often difficult [49].

Pooling Methods: Pooling has been an effective strategy for reducing the complexity of video representations and making them amenable to learning techniques. To this end, temporal pooling schemes have been proposed, such as 3D spatio-temporal gradients [36] and STIP features [42]. More recently, rank pooling has been proposed as an effective way for encoding the temporal evolution of actions (see, for example, Fernando et al. [20], Wang et al. [73], Cherian et al. [12]), however requires solving an order-constrained quadratic objective. In Wang et al. [74], a trajectory constrained deep feature pooling is proposed that pools features along motion trajectories. Several other CNN-based first-order temporal pooling schemes are proposed in [35, 83].

Our presented correlation pooling scheme is most similar to the second-order pooling approaches proposed in [7, 29] that also generates symmetric positive definite representations, but for the task of semantic segmentation of images. However, this scheme is applied on image features (such as SIFT) and cannot be easily extended to high-dimensional features generated by deep learning frameworks. In contrast, we use the frame-level prediction vectors, and the size of our correlation matrix scales by the number of action classes, which is usually much smaller than the feature dimensionality. However, to deal with higher dimensional features, we also propose a block-diagonal correlation matrix approximation, which to the best of our knowledge, is novel. Our method is also different from the Riemannian geometric approaches to action recognition proposed in Guo et al. [24], Yuan et al. [82] that uses hand-crafted image features to generate covariance descriptors.

In Cherian et al. [13], Koniusz et al. [37], we briefly touch upon the idea of higher-order pooling of CNN features for action recognition, in which we explore second-order pooling as well. However, the main focus of that paper was on third-order pooling, which further requires techniques such as kernel linearization for generating descriptors of reasonable size. In contrast, in this paper we specifically explore second-order descriptors and their variants.

Fine-grained Recognition: Early approaches to fine-grained recognition [54, 56, 57] have been direct extensions of schemes described above. Extracting mid-level appearance features, such as human body pose and motions of body-parts, have been popular for recognizing human actions [55, 70, 80, 79, 78, 86, 56]). While, there have been notable advancements in human pose estimation via deep learning methods [10, 66, 67, 76, 46], most of these models are compu-

tationally expensive and thus difficult to scale to millions of video frames that typically the datasets encompass. Moreover, most of these algorithms do not deal with occluding body-parts, which are common in long activity sequences, thus making pose-based approaches less effective. In Chéron et al. [14], human pose is used as prior to select regions of interest, and then tuning a two-stream CNN model to these regions for action recognition. While, we do not use human pose, our SMAID representation can automatically find interesting regions with significantly less computational expense. Other approaches to fine-grained action recognition include hierarchical multi-granularity action representations such as those depicted in Tang et al. [65], Lan et al. [40], Le et al. [43], grammar based models, such as Pirsiavash and Ramanan [53], Ryoo and Aggarwal [58], and schemes that first localize actions in a video and then detect them, such as Duchenne et al. [17], Bojanowski et al. [5]. In contrast to these schemes, we use the correlations between frame-level classifier predictions to get a holistic video representation.

Another popular approach to fine-grained action recognition has been via modeling human-object interactions. An object proposal framework is presented in Zhou et al. [85], that is used to produce candidate regions containing human-object interactions, from which mid-level features based on Fisher vectors are extracted for recognizing actions. A multiscale approach is presented in Ni et al. [47] that tracks the interactions between the hand and the objects in the scene explicitly via a detection-tracking framework. A similar framework for tracking people and objects via Hough forests is proposed in Gall et al. [22]. The problem has also been explored using depth cameras in Lei et al. [44], Wu et al. [77]. While, recognizing objects is useful for recognizing actions, frequently the objects being acted upon are occluded or might not have any discriminative features.

SMAID Image Representation: The proposed video sequence summarization technique (discussed in Section 4) has similarities to several prior methods. Specifically, similar to SMAID, there is motion history images (MHI) (Davis and Bobick [15]) that encodes time using image intensity (recent frames are brighter), and uses binary motion masks, thus loses texture of moving parts. SMAID uses separate image channels to capture temporal evolution. As a result, texture details of moving parts are approximately preserved per channel, while also capturing action evolution across channels. Our scheme is also different from Blank et al. [4] that uses space-time volumes as shapes for recognition. More recently, Sun and Nevatia [64] and Wang et al. [75] also propose to use image differences as inputs for training CNN models; however uses only a stack of single frame differences, where as SMAID uses the sum of absolute differences of several frames per channel (typically 7-10 frames), thereby capturing a longer temporal window.

3 Proposed Scheme

We first state our mathematical notations, followed by formally defining the activity recognition problem and our temporal correlation pooling scheme in Section 3.3. This precedes an investigation into extension of this setup for higher-order pooling (Section 3.4) and block-diagonal approximations (Section 3.5). We introduce our SMAID frame-set representation in Section 4. Next, we introduce our action classification framework using Riemannian geometry in (Section 5) and propose our end-to-end learnable three-stream CNN architecture (Section 6).

3.1 Notations

We use upper-case variables (e.g., X) for matrices (unless defined otherwise), bold-font lower-case (\mathbf{x}) for vectors, and lower-case (x) for scalars. We use δ_+^p to denote the space of $p \times p$ symmetric positive semi-definite matrices, and δ_{++}^p to denote the same for positive definite matrices. Further, $[n]$ stands for the set $\{1, 2, \dots, n\}$.

3.2 Problem Formulation

Let $\mathcal{S} = \{S_1, S_2, \dots, S_N\}$ denote a set of N video sequences, where each S_i belongs to one of M action classes with labels $\mathcal{L} = \{\ell_1, \ell_2, \dots, \ell_M\}$. Let $S = \langle f_1, f_2, \dots, f_n \rangle$, represent frames from some sequence $S \in \mathcal{S}$, and $\mathcal{F} = \bigcup_{S \in \mathcal{S}} \{f_i \mid f_i \in S\}$ be the set of all frames. Our goal is to learn a function that maps any given sequence to its correct class. To this end, suppose we have trained classifiers for each action class using a training sequence set. However, we assume it is practically difficult to train these classifiers on the sequences as a whole. Instead, the classifiers have been trained on individual frames. Let $p_m : \mathcal{F} \rightarrow [0, 1]$ be such a classifier trained to produce a confidence score for a frame to belong to the m -th action class. Unfortunately, since a single frame may not be representative of the sequence, the classifier p_m may be inaccurate at determining the action at the sequence level. As described earlier, our goals in this paper are (i) to pool the predictions of all the classifiers from all the frames in a sequence to generate a descriptor on which sequence-level action classifiers can be trained, and (ii) to improve the confidence of each classifier p_m for $m = 1, 2, \dots, M$, in making frame-level action predictions. In the sequel, we explore both these ideas.

3.3 Temporal Correlation Pooling

Using the notation defined above, let $f_1, f_2, \dots, f_n \in S$ denote a sequence of random variables corresponding to each

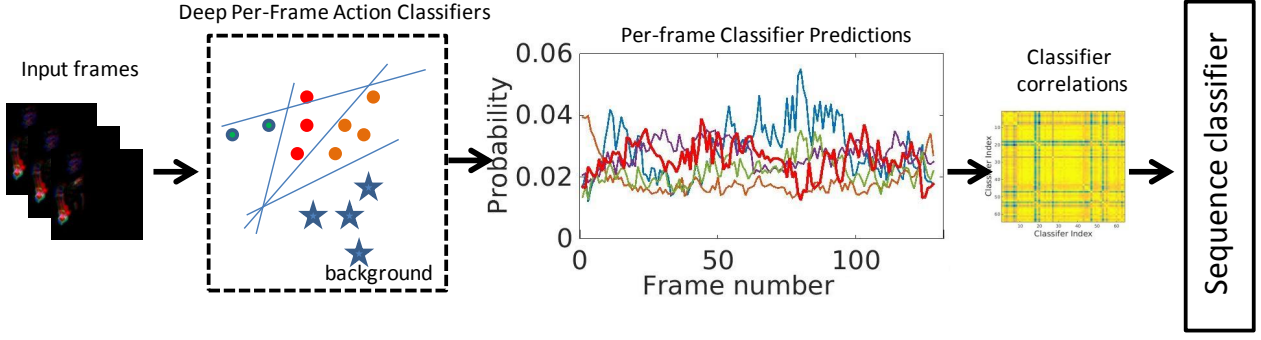


Fig. 1 Internals of our correlation pooling scheme. Each video frame passes through a pre-trained set of classifiers, and their classification scores are extracted. The temporal evolution of these scores (third block above) are pooled via our correlation scheme to generate our TCP descriptor, which is then used as the action descriptor for the video.

frame and let $p_m(f_i)$ denote the confidence that a classifier trained for the m -th action class predicts f_i to belong to class ℓ_m . Further, we assume that the scores $p_m(f_i)$ are normalized, i.e., $\sum_{m=1}^M p_m(f_i) = 1, \forall i \in [n]$. Let $\alpha^m = [\alpha_1^m, \alpha_2^m, \dots, \alpha_n^m]^T$ be a given vector of weights, where each $\alpha_i^m \geq 0$ and $\sum_{i=1}^n \alpha_i^m = 1$. Then

$$\mathbf{t}_m = [\alpha_1^m p_m(f_1), \alpha_2^m p_m(f_2), \dots, \alpha_n^m p_m(f_n)] \quad (1)$$

$$\triangleq \alpha^m \circ p_m(S), \quad (2)$$

denotes the temporal evolution of the weighted confidence of the m -th classifier for the frames in the sequence S . We call \mathbf{t}_m a *feature trajectory*. The weights α give different priority to the classifier confidences across time, and is useful when there exists prior information that certain actions happen mostly at some specific regions of a sequence (e.g., middle/beginning/end). We define our *temporal correlation pooling* action descriptor as $\text{TCP} : \mathbb{R}^{M \times n} \times \mathbb{R}^{M \times n} \rightarrow \delta_+^M$, the jk -th entry of which is given by:

$$\text{TCP}(\mathbf{t}_j, \mathbf{t}_k) = \sum_{i=1}^n \alpha_i^j \alpha_i^k p_j(f_i) p_k(f_i) = \mathbf{t}_j^T \mathbf{t}_k, \quad (3)$$

and captures the similarity between two such feature trajectories \mathbf{t}_j and \mathbf{t}_k from classifiers p_j and p_k respectively. It is clear that such a similarity computes the co-activations of the classifier scores over the sequence, and thus the co-occurrences of various activities. If $T \in \mathbb{R}^{M \times n}$ is a matrix whose m -th row is \mathbf{t}_m , then taking into account the auto-correlation nature of TCP, we also define $\text{TCP}(T)$ in matrix form as:

$$\text{TCP}(T) = TT^T \in \delta_+^M, \quad (4)$$

where δ_+^M is the space of $M \times M$ symmetric positive semi-definite matrices. Note that, we do not center each \mathbf{t}_m to the mean, as is typically done when computing correlation

matrices. As a result, the m -th diagonal entry of $\text{TCP}(T)$ is given by:

$$\text{TCP}(\mathbf{t}_m, \mathbf{t}_m) = \sum_{i=1}^n (\alpha_i^m)^2 p_m^2(f_i) \leq \sum_{i=1}^n \alpha_i^m p_m(f_i), \quad (5)$$

which is the average of classifier scores, (when α_i 's are one, it is the popular *average pooling* scheme). Thus, in essence the diagonal entries of TCP captures a lower bound to the first-order statistics. In the sequel, we propose to use $\text{TCP}(T)$ as our action descriptor. Our full pipeline is depicted in Figure 1.

The basic TCP scheme described above falls short on several fronts: (i) it only captures second-order temporal correlations, while higher-order may be more effective, (ii) the TCP matrix will be rank-deficient if the number of frames is less than the number of action classes (which poses difficulties when using Riemannian geometric methods on them [52]), and (iii) the size of TCP is quadratic in the number of classes, thus scaling them to large feature vectors may be difficult. We investigate each of these issues in detail below, thereby improving the representational power of the basic TCP scheme.

3.4 Kernelized Correlation Pooling

From (3), it is easy to see that TCP is a symmetric positive semi-definite matrix produced by an inner product between feature trajectories \mathbf{t} . It is well-known that using non-linear feature maps may better capture the complex dependencies in data, leading to superior performance [69]. To this end, we propose to embed the TCP linear inner products into a reproducing kernel Hilbert space (RKHS) via the kernel trick. Mathematically, we rewrite TCP in (3) to kernelized correlation pooling (KCP), where

$$\text{KCP}(\mathbf{t}_j, \mathbf{t}_k) = \sum_{i=1}^n \psi \left(\alpha_i^j p_j(f_i) - \alpha_i^k p_k(f_i) \right), \quad (6)$$

where $\psi(x-y)$ is a suitable non-linear positive definite function. Such a reformulation brings possibilities of incorporating rich non-linearities to capture the similarities between feature trajectories. In the sequel, we use the RBF kernel

$$\psi(x-y) = \exp(-\gamma \|x-y\|_2^2), \quad (7)$$

for a suitable choice of the bandwidth parameter γ .

As is well-known, using an RBF kernel to capture the similarity provides infinite support to KCP, thus easily avoiding it from becoming semi-definite. Further, the RBF function, with its infinite series expansion, directly captures higher-order temporal similarities¹. The bandwidth parameter γ controls the degree of smoothness of the feature trajectories. Another advantage with such a kernel embedding is the possibility to incorporate prior knowledge. For example, suppose, along with a KCP kernel $K_1 \in \delta_{++}^M$, we have another prior kernel embedding $K_2 \in \delta_{++}^M$ that captures the co-occurrence statistics of the classifiers (for example, a multi-action prior, similarity between poses across classes, etc.). Then, we can easily enhance K_1 using K_2 as $K = K_1^\alpha K_2^\beta$ or $K = \alpha_1 K_1 + \beta_2 K_2$, which are sum and product kernels respectively, for non-negative scalars α and β . We could use any number of such prior kernels, and given that the sum or product of symmetric positive definite (SPD) matrices is SPD, the resulting kernel is a valid Mercer kernel, and could be used as an action descriptor in our framework.

3.5 Block-Diagonal Kernelized Correlation Pooling

While, the above discussion assumed KCP is built on classifier scores, in this section, we extend it to work with any sequence of temporal features. Unfortunately, such an extension is not straightforward, because the size of KCP is quadratic in the feature size. Say for example, for a typical action recognition dataset, if we use the output of the last fully-connected FC8 layer (assuming a VGG/Alexnet model), the number of action classes is say 101 (used in UCF101), and thus KCP is 5151-dimensional (ignoring SPD symmetry). However, extending this setup to use intermediate layer features say from FC6 or FC7, which are 4096-dimensional, will result in KCP of size about 8 million dimensions, posing storage and computational difficulties. In this section, we propose a simple workaround for this problem via a KCP approximation, termed block-diagonal kernelized correlation pooling (BKCP).

¹ We note that in Cherian et al. [13], we proposed yet another way to incorporate higher-order pooling using higher-order outer-products, which as the experiments in that paper show, is complementary to using KCP directly. Specifically, KCP captures the higher-order co-occurrences of the CNN features in the temporal mode, while the scheme in [13] captures the occurrences of multiple CNN features across classifiers.

In a nutshell, our main idea of the BKCP approximation is to reduce a full KCP matrix computed over all the feature dimensions into a block-diagonalized KCP, where each diagonal block of KCP captures the second-order correlations between only a subset of the features. Given that we could treat each block-diagonal of such a matrix as an independent KCP, we could scale the size of BKCP linearly in the feature size. On the downside, we ignore some correlations that could be important. To accommodate this, we repeat this BKCP construction process several times after randomly permuting the feature indexes. Such a scheme is reminiscent of the popular product quantization techniques [32] and model averaging schemes [27].

Mathematically, suppose $\theta(f_i) \in \mathbb{R}^d, i \in [n]$ represents features from some layer of a CNN for frames f_i . Further, let $\Theta \in \mathbb{R}^{d \times n}$ be a feature trajectory matrix built on $\theta(f_i)$ such that the k -th row Θ_k is given by $\theta_k(f_i), i = 1, 2, \dots, n$, which is the k -th feature trajectory. BKCP aims to quantize each $\theta(f_i)$ into the Cartesian product of several smaller features (in distinct sub-dimensions) and then compute the kernelized correlation matrix on such sub-vectors. That is, let $\theta(f_i) \in \mathbb{R}^p \times \mathbb{R}^p \times \dots \times \mathbb{R}^p$ times. Suppose, we randomly (jointly) permute the dimensions of the feature trajectories using a permutation matrix $\pi \in \Pi$, where Π is the set of all $d \times d$ permutation matrices², and we denote this shuffling as $\pi \circ \theta(f_i) = \pi(\theta)(f_i)$. Let $\pi(\theta)(f_i)^{(p(k-1)+1:p(k-1)+p)} \in \mathbb{R}^p$, ($k \in [d/p]$), denote such a sub-vector of p dimensions of $\pi(\theta)(f_i)$ starting at dimension $p(k-1)+1$. Then, $\forall u, v \in [p(k-1)+1, p(k-1)+p]$, we define the Block KCP (BKCP) approximation to KCP for the block at $(p(k-1)+1, p(k-1)+1)$ extending to (pk, pk) in KCP as:

$$\text{BKCP}_\pi^{p(k-1)+1:pk}(\Theta_u, \Theta_v) = \sum_{i=1}^n \exp\left(-\gamma \|\pi(\theta)_u(f_i) - \pi(\theta)_v(f_i)\|_2^2\right), \quad (8)$$

where we have substituted the RBF kernel for KCP as described in the last section and the notation $\pi(\theta)_u(f_i)$ denote the ui -th entry of Θ after permuting its rows by π . We extend this definition to cover all such permutations of dimensions, and we define the approximation to KCP as:

$$\text{BKCP}^{p(k-1)+1:pk}(\Theta_u, \Theta_v) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \text{BKCP}_\pi^{p(k-1)+1:pk}(\Theta_u, \Theta_v). \quad (9)$$

In detail, the steps for constructing BKCP descriptors are as follows. First, we select a permutation $\pi \in \Pi$, and permute all the rows of the feature trajectory matrix Θ using π . Then, we compute KCP on each disjoint set of p -dimensional blocks (sub-vectors or rows of the permuted Θ). For example, one

² Note that such shuffling can be implemented in linear time; we do not need to generate such large permutation matrices explicitly.

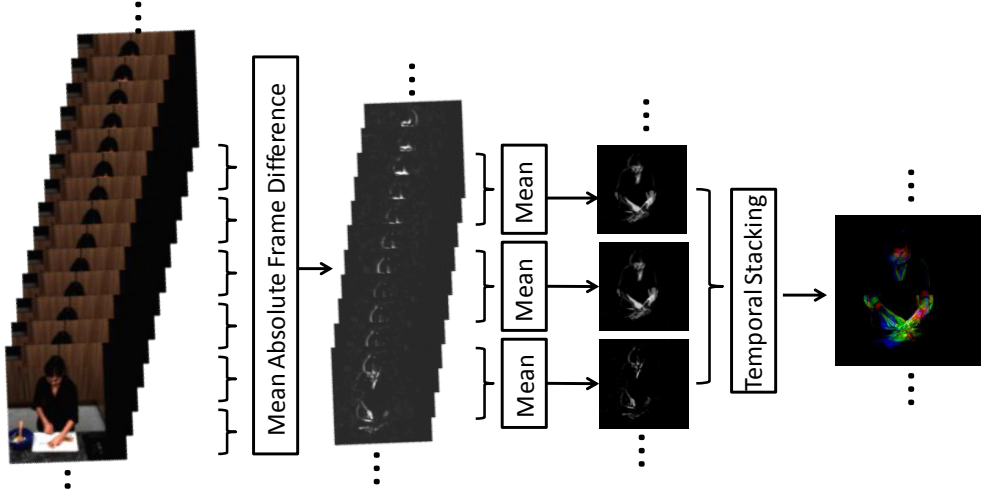


Fig. 2 Schematic illustration of the steps involved in generating our SMAID images. We first convert the frames to gray-images, which are then differenced (second frame stack). These difference images are averaged to generate MAID images (third stack), which are then stacked across channels to generate SMAID.

block may have features from dimensions 1 to p , next block from dimensions $p + 1$ to $2p$, etc. The resulting KCP will be block-diagonal (assuming the blocks are selected in order). Next, we select another permutation π from Π and repeat the process. Finally, the different KCPs from different π s are averaged to generate one approximate block-diagonal BKCP (9). Assuming d -dimensional features, KCP as defined in (6) will have a size $d(d - 1)/2$, while BKCP will have a size $d(p - 1)/2$, which for appropriately chosen and fixed p blocks scales linearly with d .

4 SMAID Image Representations

Success of any pooling scheme depends on the quality of the features (or classifier scores) used. This is because, more noise in the features (or predictions) leads to diluting the feature correlations. While, the two stream model is popular and is empirically seen to be effective, it discards the coupling between optical flow and appearance streams. For example, in [19], a fusion of intermediate CNN layers is proposed, where the pooling between flow and RGB streams are accounted for earlier than the last layer. Such a fusion synchronizes the two disparate feature maps and allows for joint inference at the last layer. In this section, we propose a much cheaper fusion scheme using differences of frames, that approximates flow and appearance.

For a sub-sequence $S_{\tau+1:\tau+\zeta} = \{f_{\tau+1}, \dots, f_{\tau+\zeta}\} \subseteq S$ containing ζ consecutive frames, we define the *mean absolute image difference* (MAID) representation of $S_{\tau+1:\tau+\zeta}$ as:

$$\text{MAID}(S_{\tau+1:\tau+\zeta}) = \frac{1}{\zeta - 1} \sum_{j=2}^{\zeta} |f_{\tau+j} - f_{\tau+j-1}|. \quad (10)$$

As is clear, such a representation aggregates small motions over ζ consecutive frames and summarizes them in a single object with the same dimensionality as a single frame. However, such a representation loses the long-term temporal evolution of actions; to circumvent this we stack several such MAID images corresponding to consecutive non-overlapping sub-sequences as separate image channels. That is, suppose $S' = S_{\tau+1:\tau+\beta\zeta}$ is a subsequence of S containing $\beta\zeta$ frames. Then, we define our Stacked MAID (SMAID) representation as:

$$\text{SMAID}(S') = \bigotimes_{j=1}^{\beta} \text{MAID}(S_{\tau+(j-1)\zeta+1:\tau+j\zeta}), \quad (11)$$

where the operator \bigotimes represents stacking MAID images into the third mode of a 3D tensor. To restrict the SMAID cross-channels to only allow temporal evolution of the actions, we reduce the original color images to gray-scale MAID images before stacking them. The overall SMAID pipeline is depicted in Figure 2. See Figure 4 and Figure 7 for more SMAID illustrations.

Next, this SMAID image representation is fed to a three-stream CNN; consisting of separate streams for appearance, flow, and SMAID frames. Due to the demonstrated performance benefits, we chose a 16-layer VGG network [9], pre-trained on the Imagenet dataset, to form the CNN classifiers for the individual data streams. A schematic illustration of our full pipeline is depicted in Figure 3.³

³ As we fine-tune the VGG network from a pre-trained ImageNet model, we use $\beta = 3$ for SMAID in our implementation.

5 Classification on the Riemannian Manifold

Now that, we have provided all the details for generating a second-order action descriptor for a given video sequence, let us move on to algorithms for classifying SPD matrices in an SVM setup. Our overall classification pipeline is depicted in Figure 3. As is clear, the kernelized correlation matrices are symmetric positive definite (SPD) objects themselves; each sequence generating one such object. It is well-known that these matrices belong to the strict interior of the cone of positive semi-definite (PSD) matrices. While, PSD can be treated as objects in the Euclidean space under the natural Frobenius norm, it is often found that resorting to a non-linear geometry on SPD matrices can avoid unlikely or impossible outcomes (such as for example, nearest neighbors to an SPD matrix is restricted to be only SPD matrices, instead of PSD), thereby improving application performance [52, 1]. Typically, this non-linear geometry is imposed via the respective similarity measure used to compare SPD matrices. Among the commonly used such measures [52, 11, 1], we will be exploring two, namely (i) the log-Euclidean metric [1] and (ii) the Jensen-Bregman logdet divergence [11], as they are known to induce valid Mercer kernels on SPD matrices. We detail each of these measures and their respective kernels below.

5.1 Log-Euclidean Metric

For two KCPs $C_1, C_2 \in \delta_{++}^d$, the log-Euclidean distance between them is given by:

$$\text{dist}_{LE}(C_1, C_2) = \|\text{Log } C_1 - \text{Log } C_2\|_F, \quad (12)$$

where Log is the matrix logarithm, which makes an isomorphic mapping between an SPD matrix C and a symmetric matrix $\text{Log } C$, the latter uses the Euclidean geometry and thus similarity could be computed using the standard Frobenius norm. An advantage of using dist_{LE} is that it decouples the constituent matrices, such that the Log operator could be applied during data pre-processing, after which evaluating the similarity involves only computing Euclidean distances, which can be done very fast. However, gradient computations using this measure is difficult [1], which makes end-to-end learning difficult. An RBF kernel using the log-Euclidean metric for SVM classification is introduced in [45] and has the following form:

$$K_{LE}(C_1, C_2) = \exp(-\xi \text{dist}_{LE}(C_1, C_2)^2), \quad (13)$$

where ξ is a bandwidth parameter. Note that, the log-Euclidean kernel can be looked at as the limit of the popular power-normalization strategy, which is known to combat *burstiness* [31], i.e., certain classifiers firing more frequently than others. In addition, the log-Euclidean kernel can be directly

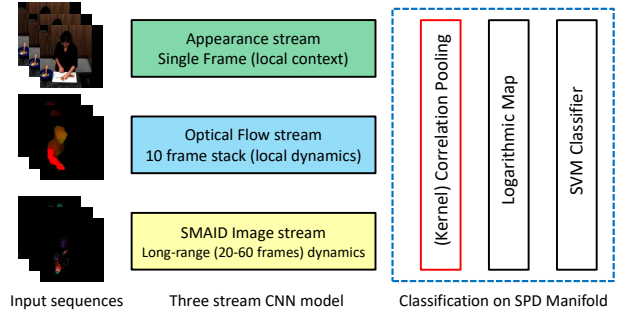


Fig. 3 An illustration of our overall CNN architecture and our pooling scheme. We use a non-linear feature pooling scheme based on Riemannian geometry to generate an action descriptor.

applied to each block of our BKCP descriptor separately, thus making the scheme efficient (as otherwise one needs to compute the singular values of a very large KCP matrix).

5.2 Jensen-Bregman Log-Det Divergence

Another popular similarity measure on SPD matrices is the recent Jensen-Bregman Log-Det divergence (JBLD) [11] (also called Stein divergence [62]), which for two KCPs C_1 and C_2 has the following form:

$$\text{dist}_S(C_1, C_2) = \log \det \left(\frac{C_1 + C_2}{2} \right) - \frac{1}{2} \log \det (C_1 C_2). \quad (14)$$

In contrast to the log-Euclidean metric, JBLD is not a Riemannian measure, instead is a symmetric Bregman divergence which captures the information divergence between a function and its first-order Taylor approximation (the function is $-\log \det$ in this case). It is related to the Bhattacharya distance [30] between two zero mean Gaussian distributions with covariances C_1 and C_2 . In contrast to the log-Euclidean metric that needs to compute the matrix logarithm of the constituent matrices, JBLD needs only the matrix determinant, which is computationally cheaper. In [62], a kernel is defined using JBLD as defined below:

$$K_S(C_1, C_2) = \exp(-\xi \text{dist}_S(C_1, C_2)), \quad (15)$$

$$\forall \xi \in \left\{ \frac{k}{2}, k = 1, \dots, d-1 \right\} \cup [d, \infty),$$

where the bandwidth parameter ξ is defined only for certain values. In contrast to the log-Euclidean metric, JBLD offers computationally cheaper gradients, as will be explored in the next section.

6 End-to-End CNN Training

In this section, we explore an end-to-end CNN architecture that learns the action descriptors and the classifiers jointly

via gradient back-propagation. As is the case with any end-to-end CNN models, the main challenge in designing this model is to define the gradients of the objective with respect to the inputs. There have been several previous attempts at implementing end-to-end second-order CNN models. In [29], the log-Euclidean metric is used to define the CNN loss function. While taking gradients of this metric is challenging, it also leads to a flattening of the matrix, leading to very large fully-connected layers that scales quadratically with the number of data classes. In Huang and Van Gool [28], a CNN model that takes SPD matrices as input is presented. Another recent attempt (Yu and Salzmann [81]) is to map the second-order SPD matrices into a lower-dimensional SPD manifold through parametric second-order transformation, followed by parametric vectorization. However, such parametric transforms also introduce additional capacity to the networks that needs to be learned. In contrast to all these methods, we propose to directly use second-order similarity measures to define loss functions, which as we show below leads to simple and straightforward gradient formulations, without the need for introducing any new parameters into the framework. We explore two such loss functions, namely (i) using the Jensen-Bregman Logdet Divergence as introduced in (14), and (ii) using the simple Frobenius norm.

6.1 End-to-End Learning Using Stein Divergence

Suppose $T \in \mathbb{R}^{M \times n}$ denotes the CNN feature trajectories⁴ (from say the FC8 layer of a standard VGG/ Alexnet model) for n frames in a sequence and M action classes. Further, let Y denote an $M \times M$ diagonal ground-truth label matrix for a ground-truth label ℓ associated with T ; the jj -th diagonal entry of Y is defined as

$$Y_{jj} = \begin{cases} 1/(1 + (M-1)\epsilon), & \text{if } j = \ell_i \\ \epsilon/(1 + (M-1)\epsilon), & \forall j \neq \ell_i \end{cases} \quad (16)$$

where we assume ϵ is a small number (say 10^{-5} used in our experiments). This encoding of ground truth class label is similar to the standard one-off encoding used with a softmax cross-entropy loss framework. However, given that we propose to use similarity measures defined on SPD matrices in our loss, we cast the label in a matrix form and use a small ϵ regularization to make sure this matrix SPD.

Suppose, we have a training set consisting of such sequences of CNN feature trajectories $\mathcal{T} = \{T^1, T^2, \dots, T^N\}$ for video sequences in \mathcal{S} and their associated ground-truth encoded matrices $\mathcal{Y} = \{Y^1, Y^2, \dots, Y^N\}$. Then using the JBLD measure introduced in (14), we define the TCP CNN

loss as:

$$\text{loss}(\mathcal{T}, \mathcal{Y}) := \sum_{\forall (T, Y) \in \mathcal{T} \times \mathcal{Y}} \left[\log \det \left(\frac{Y + \text{TCP}(T)}{2} \right) - \frac{1}{2} \log \det Y - \frac{1}{2} \log \det (\text{TCP}(T)) \right], \quad (17)$$

$$\text{where } \text{TCP}(T) = \frac{1}{n} T T^T.$$

For implementing back-propagation, we need the gradient of loss with respect to a data matrix T (with associated label matrix Y) and is as follows:

$$\frac{\partial \text{loss}(T, Y)}{\partial T} = \frac{2}{n} \left\{ \left(\text{TCP}(T) + Y \right)^{-1} - \frac{1}{2} \text{TCP}(T)^{-1} \right\} T. \quad (18)$$

6.2 End-to-End Learning Using Frobenius Norm

A difficulty usually encountered with the gradient defined in (18) is the need to compute the matrix inverse, which is expensive and will also sometimes lead to numerical instability. Thus, we also propose to use the matrix Frobenius norm to define the CNN loss, which completely avoids these issues. As this loss will not require the label matrix Y to be SPD, we assume $\epsilon = 0$ in this case in (16). Reusing the notations from the last section, we define the new loss as:

$$\text{loss}(\mathcal{T}, \mathcal{Y}) := \sum_{\forall (T, Y) \in \mathcal{T} \times \mathcal{Y}} \| \text{TCP}(T) - Y \|_F^2, \quad (19)$$

and the respective gradient with respect to a data matrix T has the form:

$$\frac{\partial \text{loss}(T, Y)}{\partial T} = \frac{2}{n} \left(\text{TCP}(T) - Y \right) T \quad (20)$$

Empirically, it is observed that using the softmax output of the FC8 CNN layer for constructing the above losses leads to better convergence of the models.

7 Experiments

In this section, we evaluate the usefulness of our proposed framework on four datasets. Two of these datasets, namely the MPII Cooking activities dataset [56], and the JHMDB dataset [33], are standard fine-grained benchmarks. We also provide evaluations on HMDB and UCF101 datasets, which are standard benchmarks with fine-grained as well as coarse action categories. As alluded to earlier, we use the standard 16-layer Imagenet pre-trained VGG deep learning network [9], which we fine-tune for each of the input modalities. Below, we provide details of these datasets, data preparations, evaluation protocols, and our results.

⁴ With a slight abuse of previously introduced notations, we assume T to be raw feature trajectories without any scaling or normalization.

7.1 Datasets

MPII Cooking Activities Dataset [56]: This dataset consists of high-resolution videos of cooking activities captured by a static camera. The videos are of 12 different people cooking various dishes and consists of 64 distinct activities spread across 3748 video clips and one background activity (1861 clips). There are over 800K frames and the activities range from coarse subject motions such as *moving from X to Y*, *opening refrigerator*, etc., to fine-grained actions such as *peel*, *slice*, *washing hands*, *cut ends*, *cut apart*, etc. This dataset is challenging due to several reasons, namely (i) the classes are very unbalanced – there are certain activities that have only about 1K frames over the entire dataset, (ii) there is significant intra-class variability as the participants are only asked to prepare one of a set of 14 dishes and allowed to cook in their own styles, and (iii) there are no annotations of objects in the scene, and the tools used for actions are very small (such as spice folder, knife, etc.) and thus hard to detect.

HMDB Dataset [39]: It consists of 6766 videos from 51 different action categories, mostly web videos of low resolution and quality. Each video clip is a few seconds long. There are videos which are mostly downloaded from the Internet and contains significant changes in lighting, view-points, slight camera motions, and inter/intra-class variability, that makes the recognition task challenging. The dataset includes videos that are not person centered, and action may undergo occlusions, or strong camera motions, thus making recognition very challenging.

JHMDB Dataset [33]: This dataset is a subset (960 videos) of the HMDB dataset consisting of 21 actions, but contains videos for which the human limbs can be clearly identified. The dataset contains action categories such as *brush hair*, *pick*, *pour*, *push*, etc., in low resolution videos.

UCF101 Dataset [61]: This dataset contains 13320 videos spread in 101 action categories. This dataset is different from the above ones in that it contains mostly coarse sports activities with strong camera motion and low resolution videos.

Evaluation Protocols: Following the standard protocols, we use mean average precision over 7-fold cross-validation on the MPII dataset. Other datasets use mean average accuracy on 3-splits. For the former, we use the evaluation code published with the dataset.

7.2 Preprocessing

As the original MPII cooking videos are very high resolution, while the activities happen only at certain parts of the scene, we used morphological operations to estimate a

window of the scene to localize the action. The videos are then cropped to these regions, resized to 224×224 size and used to train the VGG networks. Precisely, for every sequence, we first convert the frames to half their sizes, followed by frame-differencing, dilation, smoothing, and connected component analysis. This results in a binary image for every frame, which are then combined across the sequence and a binary mask is generated for the entire sequence. We use the largest bounding box containing all such connected components in this binary mask as the region of the action, and crops the video to this box. To compute optical flow, we used the TVL1 OpenCV implementation, thresholding flow to ± 20 pixels, rescaled to 0–255, and saved as a JPEG image for storage efficiency as described in [60].

7.3 SMAID Image Parameters

As noted earlier, SMAID images summarize long-range actions into a compact image representation. There are two parameters for this representation: (i) number of frames that can be effectively summarized in a SMAID channel (ζ in (10)), and (ii) number of channels that can be stacked to capture the dynamics (β in (11)). Depending on the sequences, too many frame differences for (i) might result in a cluttered image that may not be useful for learning actions, while too less frames might lead to very sparse images. For (ii), while a 3-channel stack will render the SMAID as equivalent of an RGB image and thus RGB based CNN architectures could be used, higher-number of channels will require redesigning the network, and also leading to more CNN parameters. See Figure 4 for example frames from the UCF101 dataset for various number of frames encoded per channel in a 3-channel SMAID setup.

To understand the effect of these parameters, we progressively increased (i) and (ii) on a subset of the UCF101 split-1 training set containing videos that had limited camera motion, and evaluated on a small validation subset. All the plots are generated on an Alexnet network trained from scratch on SMAID images (so that they are compared against the same baselines). In Figure 5, we plot the classification accuracy. On the one hand, higher number of frames per channel in SMAID leads to performance improvements, but with more than a certain number (seven), the performance drops, perhaps because of increasing clutter (see Figure 4). On the other hand, with increasing number of SMAID channels (beyond three), the performance is seen to decrease, which is surprising. We think this behavior is perhaps because of the typical network architecture that we use (Alexnet), which is designed for RGB images and thus is inadequate for a SMAID image with more than three channels. Thus, in the sequel, we use a 3-channel SMAID stack, with 15 frames per channel for HMDB and UCF101 datasets. However, we found a 7 frames per channel works best for MPII cooking

Experiment	MPII-mAP (%)	JHMDB-Avg.Acc.(%)	HMDB-Avg. Acc.(%)	UCF101-Avg. Acc. (%)
RGB	33.9	51.5	40.9	82.0
FLOW	37.6	54.8	47.5	85.1
SMAID	35.4	61.1	41.1	72.1
RF	38.1	55.9	53.6	88.5
RFS	39.5	62.6	54.4	88.8

Table 1 Sequence level comparison (using average pooling) on MPII cooking dataset, JHMDB, HMDB (split1) and UCF101 (split1).

Expt	MPII CP	MPII KCP	MPII BKCP	JHMDB CP	JHMDB KCP	JHMDB BKCP	HMDB CP	HMDB KCP	HMDB BKCP	UCF101 CP	UCF101 KCP	UCF101 BKCP
RGB	49.7	52.7	55.2	47.6	52.3	39.2	52.8	56.7	58.7	79.1	82.2	76.9
FLOW	55.6	60.6	61.4	57.9	60.4	64.9	45.9	53.3	57.2	83.1	86.1	83.4
SMAID	51.3	55.7	59.6	50.2	64.6	55.6	49.4	52.9	52.1	74.2	71.7	70.7
RF	60.0	64.4	65.6	59.4	63.4	61.5	57.1	65.2	68.1	86.2	87.8	87.5
RFS	62.1	66.1	68.0	62.0	72.7	69.4	57.8	66.7	68.5	87.2	88.3	87.9

Table 2 Comparison of correlation pooling (TCP) against kernelized correlation pooling (KCP) on FC8 layers and block-diagonal kernelized correlation pooling (BKCP) on FC6 CNN layer on MPII, JHMDB, HMDB (split1), and UCF101(split1) datasets. We use mAP on MPII evaluations and mean classification accuracy on the other datasets.

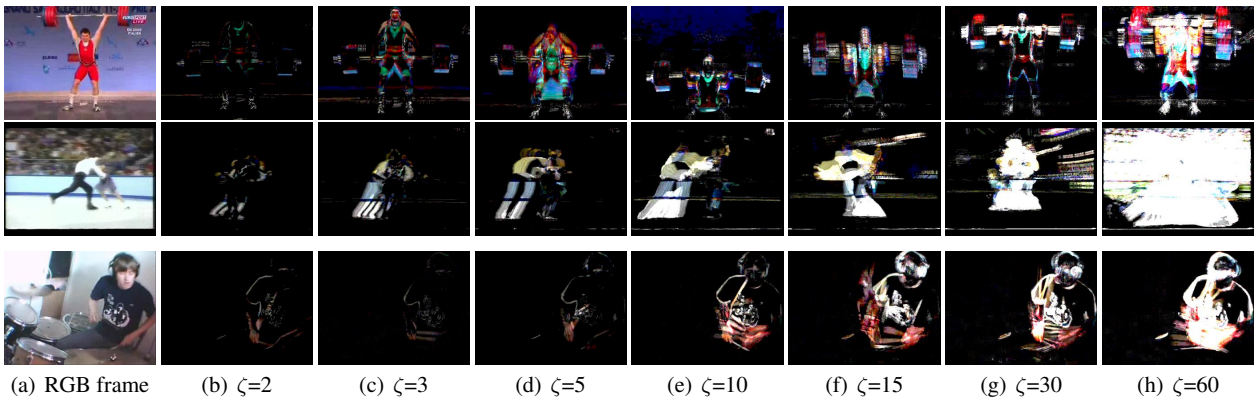


Fig. 4 Comparison of 3-channel SMAID images with varying number of frames summarized per channel (ζ) for two sequences from UCF101 dataset. As is clear, higher ζ leads to cluttered image, while smaller ζ fails to capture sufficient motion. Also, note that for each SMAID image, the temporal order is mapped to colors Red < Green < Blue.

activities and JHMDB datasets. With these configurations, each SMAID image captures subsequences of 45 frames in UCF101 and HMDB-51, and uses 21 frames in JHMDB and MPII datasets.

We would also like to point out that SMAID with only one frame-difference per channel is equivalent to some of the recent proposals described in [75] and [64]. However, as is clear from Figure 5(b), more frames per channel is significantly better. Further, looking back at Figure 5(a), a single channel SMAID is a grayscale image, similar to a motion history image [15]. However, using more channels is clearly beneficial. These two plots substantiates that the design of SMAID is better than existing frame summarization techniques based on frame differencing.

7.4 Parameters for BKCP

The Block-diagonal approximation for KCP has two parameters, namely (i) the length of the subvectors (p defined in (9))

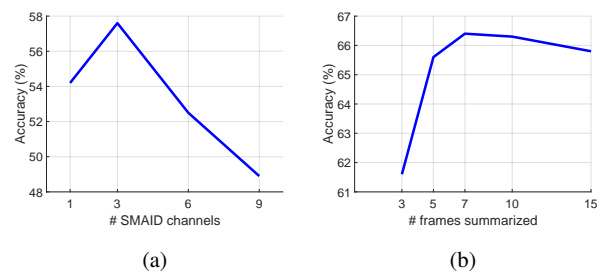


Fig. 5 Evaluation of the effect of increasing number of SMAID channels (keeping number of frames per channel fixed at 5) and increasing number of frames per channel (while keeping the number of channels fixed at 3) on UCF101 split1 using Alexnet.

and the number of feature permutations to be tried to estimate the BKCP descriptor. For the former, as is clear, higher values of p demands higher computations, while lower p will ignore important correlations; in the limit $p = 1$ is only the

diagonal correlation matrix, which is not useful for pooling. In Figure 6, we evaluate the advantages of using BCKP on FC6 features from a VGG-16 CNN model. These are 4096-dimensional features. We used a linear SVM on the resulting block-diagonalized pooling features after embedding them in the Euclidean space using the log-Euclidean projection (described in Section 5.1).

The plot in Figure 6(a) shows that higher dimensionality of the sub-vectors is not beneficial for classification. This is perhaps due to the fact that such high dimensions result in mostly ill-conditioned blocks in KCP, which when passed through the log-operator will amplify the noise contained than useful action signals, thereby reducing their benefits. We found that using sub-vectors of length $p = 16$ showed the best performance. In Figure 6(b), we investigated the influence of increasing the size of the permutation set. Given that, the cardinality of such a set increases factorially with the number of feature dimensions, estimating a reasonable size is important. Surprisingly, we found that a small number of such permutations is sufficient to get a reasonable accuracy. As motivated by this plot, we use three random (but fixed fixed for the entire dataset) permutation matrices for model averaging.

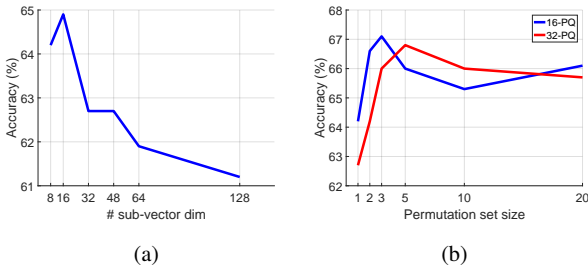


Fig. 6 Left: Evaluation of the effect increasing the number of BCKP sub-vector dimensions. Right: Evaluation of averaging BCKP over various dimensional permutation set sizes.

7.5 Experiment Setup

All the three CNN streams (RGB, Flow, and SMAID) are trained separately. Among the end-to-end CNN loss variants (Frobenius norm versus Stein divergence), we use the Frobenius norm due to its superior speed and numerical stability. We found that the performance of Frobenius norm is very similar to the standard softmax cross-entropy loss. We use sub-sequences of 30 frames for computing the correlation matrices in the end-to-end setup. Given a fixed CNN batch size (number of frames), we could not use more frames per sequence, as this limits the number of sequences that could be used in a training batch, and thus restricting the batch diversity (different action classes in the same batch). Less di-

verse such batches are known to impact convergence. Once the CNNs are trained, we use a forward pass to compute per-frame features, which is then used to generate sequence level TCP descriptors and variants. These descriptors are then used in a Riemannian geometry based SVM classification framework, thus utilizing the power of non-linear geometry. We found that this provided significantly better accuracy than just using the end-to-end learned model.

Specifically, in all the experiments to follow, we use the following settings. We use a VGG-16 model pre-trained on UCF101 dataset from [19] to fine-tune the models for JHMDB, MPII Cooking activities, and the HMDB-51 datasets. As alluded to above, we use single RGB images for the RGB stream, a stack of ten consecutive optical flow images for the flow stream, and three-channel 21–45 frame compressed SMAID images for the respective CNN stream. To train the SMAID CNN stream, we use the RGB stream of the above pre-trained model for initialization of the stream weights, which seemed to perform significantly better than learning from scratch. For fine-tuning, we used a fixed learning rate of 10^{-4} and a momentum of 0.9. We used the Caffe toolbox⁵ for our CNN implementations. We also applied the standard data augmentation techniques (such as mirroring) on the data inputs. For the RGB stream, the CNN iterations usually converged in about 20k iterations, the optical flow stream 40–60k iterations, and about 70k iterations for the SMAID stream.

During testing, predictions from each of the three streams (output of FC8 layer, which is normalized to be in $[0, 1]$ after subtracting the minimum value) are aggregated at the sequence level, kernelized (using a $\gamma = 1$) and later vectorized after taking the matrix logarithm. For the MPII dataset, we used the provided training and validation sets. For JHMDB, we used 95% of the training set to fine-tune CNNs, 5% as validation. For the UCF101 dataset, we used the pre-trained CNN models from [19] for the RGB and FLOW streams. For HMDB dataset, we trained our three streams by fine-tuning those used for UCF101.

8 Results

In this section, we provide systematic evaluations of our various schemes on the four datasets. The notations RGB, FLOW, and SMAID denote the respective frame-level features. We denote the combinations of RGB+FLOW as RF and RGB+FLOW+SMAID as RFS, where the combinations are either averaged over their softmax CNN outputs for frame-level predictions, or their log-mapped features concatenated when using the correlation pooling schemes.

⁵ <http://caffe.berkeleyvision.org/>



Fig. 7 Qualitative SMAID and the associated appearance images from the JHMDB dataset (top) and the MPII Cooking activities dataset (bottom).

8.1 Evaluating the SMAID Representation:

First, we evaluate our SMAID representation at the frame-level against alternatives such as (i) using only a single stream image model RGB and (ii) using only optical flow stream FLOW. In Table 1, we provide these results on the four datasets. As is clear, SMAID improves the frame-level confidences by about 4% on the MPII and about 10% on JHMDB. However, it seems less effective on HMDB and UCF101 datasets. This is not surprising as SMAID uses frame differences, and given that there is significant camera motion in HMDB and UCF101 datasets, its influence is marginal. Further, it appears that SMAID offers complementary cues for recognition not captured by either appearance or flow or their combination; improving the accuracy from 38.1% to 39.5% on MPII sequences and by about 7% on JHMDB, which we believe is substantial, thus clearly delineating the operational characteristics of this new data modality.

8.2 Correlation Pooling:

Next, we evaluate our correlation pooling (TCP) scheme and its kernelized variants (KCP and BKCP) on CNN features (FC8 for TCP and KCP, FC6 for BKCP) from the three input modalities. The results are shown in Table 2. Comparing these results to those in Table 1, shows that KCP improves sequence level performance substantially for both MPII, JHMDB, and HMDB datasets; from 39.5% to 66.1% for RFS on MPII, from 62.6% to 72.7% on JHMDB, from 54.4% on HMDB-51 to 66.7%. However, correlation pooling seems less effective on UCF101, which we believe is due to two reasons, namely (i) there is a strong bias that per-frame context offers for UCF101 video sequences (as is similarly pointed out by [25]), as a result there is nothing

Experiment	MPII KCP mAP(%)	JHMDB KCP Avg.Acc.(%)
LE Kernel	66.1	72.7
Stein kernel	68.5	62.5

Table 3 Comparison of performance when using different kernels in SVM for classifying the kernelized correlation matrices over the three input modalities.

more in the temporal evolution of classifiers that is captured by our second-order method, and (ii) the co-occurrences of actions (or confusion between classes) is minimal and offers a very weak signal that can be taken advantage of by our scheme. Table 2 also shows that kernelizing the temporal correlations (TCP versus KCP or BKCP) is always useful; demonstrating a consistent 5% improvement from its non-kernelized variant.

8.3 KCP Classification Kernel

As reviewed in Section 5, there are popularly two SVM kernels on SPD matrices, the Stein kernel and the log-Euclidean kernel. In Table 3, we show results comparing these two kernels on the MPII Cooking Activities and the JHMDB datasets. As is clear, either kernel performs differently and generate improvements, suggesting that it is better to cross-validate each of the kernels on the respective datasets to choose the right one. Given that the improvements produced by the log-euclidean kernel on the JHMDB dataset is significantly higher than the improvements by the Stein kernel on the MPII dataset and further noting the computational advantages as described in Section 5.1, we decided to use the log-euclidean kernel in the sequel.

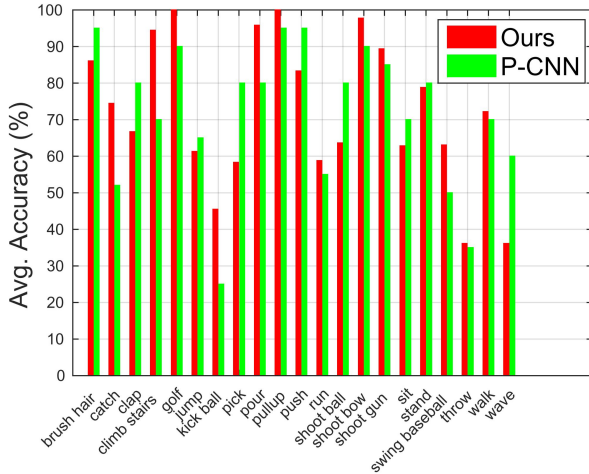


Fig. 8 KCP per-class classification accuracy of sequences in the JHMDB dataset in comparison to the in [14].

Experiment	MPII TCP mAP(%)	MPII KCP mAP (%)
RGB	44.7	48.0
FLOW	32.2	37.0
SMAID	50.1	53.3
RF	46.4	50.9
RFS	53.1	59.8

Table 4 Improvements between TCP and KCP when using an Alexnet CNN model instead of a VGG-16 network (compare to Table 2 on the MPII Cooking activities dataset).

Action	RF mAP (%)	RFS mAP (%)
Change Temperature	32.1	50.7
Dry	46.6	53.6
spice	29.4	34.9
put on cupboard	24.1	18.3

Table 5 Analysis of per-class recognition accuracy on the MPII dataset with avg. pooling when using RGB + FLOW against RGB+FLOW+SMAID.

8.4 Alternative CNN Architectures

In this subsection, we compare the performance reported in Table 2 when using an Alexnet model instead of a VGG-16 network. While, it is obvious that the performance will be inferior, our goal is to verify if the correlation pooling schemes demonstrate a similar trend. In Table 4, we show the results of this experiment. As is clear, using VGG-16 shows better results against Alexnet model in all the input modalities, and we also find that KCP is always better than KCP, thus confirming that the trend that we see is independent of the network architecture.

Action	Avg. Pool mAP (%)	KCP mAP (%)
Change Temperature	15.1	53
Dry	27.7	51
Fill water from tap	10	42
Open/close drawer	25.2	65.1

Table 6 An analysis of per-class action recognition on MPII dataset accuracy when using average pooling and KCP pooling (the top classes corrected by KCP pooling).

8.5 Comparisons to the State of the Art

Finally, in Tables 7, 8, and 9, we present comparisons of our full framework against state-of-the-art approaches on the four datasets. On the MPII cooking activities dataset, our kernelized correlation pooling scheme shows an overall mAP of 68% (Table 2). This is better than the results in recent CNN based approaches such as [14] (62.3%) and better than non-CNN based, yet state of the art schemes such as [41] (66.8%). Further, we see that incorporating trajectory features into our framework substantially improves our accuracy further to 74.7% (Table 7) outperforming all other approaches. On the JHMDB dataset, our correlation pooling scheme provides an average accuracy of 62%, while the kernelization scheme improves this to 72.7%. In comparison to the CNN based results in [14], our results are about 10% better. Further, incorporating BKCP and trajectory features increases our performance to 77.3%, which is better than the next best method by about 5.1%. These comparisons clearly demonstrate the effectiveness of our methodology against prior works. On HMDB dataset, our combination of KCP, BKCP, with dense trajectory features demonstrate state of the art performance, better by about 1.3%, on a similar capacity VGG-16 model [19], and providing competitive performance (70.3% versus 70.5% ours) against a more sophisticated Residual network based two stream model (with inter-stream coupling). Further, on the UCF101 dataset, we showcase results very close to the state of the art (92.4% against 93.5% in [19] using VGG-16, and 94.6% using ResNet-50). Given that ours scheme is independent of the CNN architecture, as is illustrated by our results in Table 4 and the consistent improvements that it produces in comparison to first-order pooling schemes (Table 1 and Table 2), already substantiates the usefulness of our second-order method.

8.6 Analysis of Results

In this section, we provide more analysis of our results, summarizing when second-order methods improved the performance in the datasets that we use. In Tables 4 and 6, we analyze the actions that are mostly corrected by our pooling scheme. As is seen, actions such as *change tempera-*

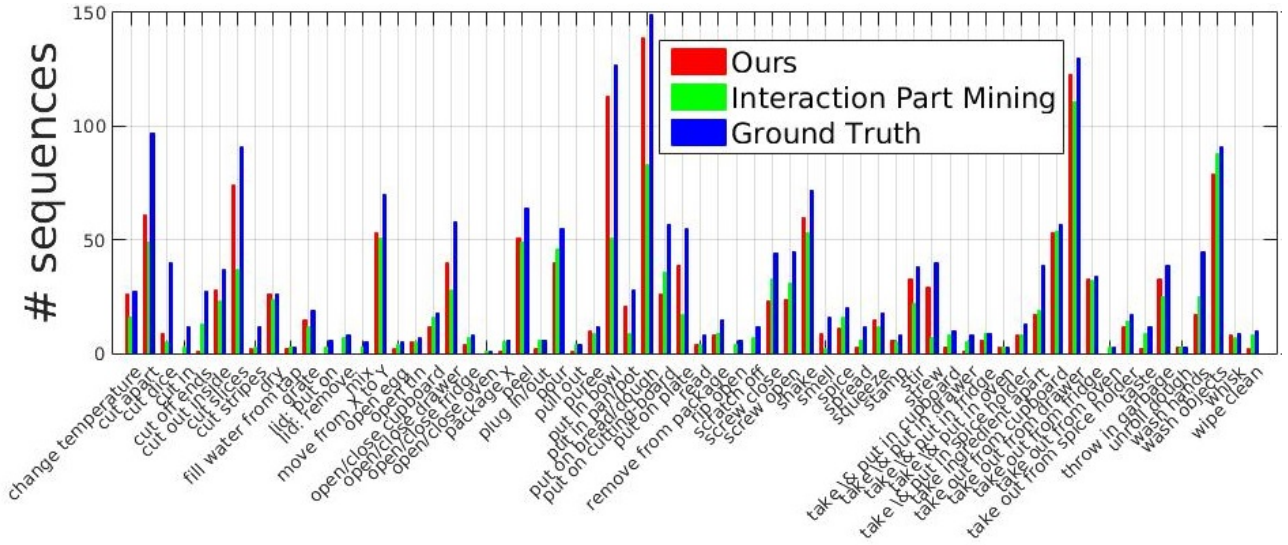


Fig. 9 KCP per-class accuracy for the MPII Cooking activities dataset against the number of ground truth videos in each class. We also compare our results to those from [85].

Algorithm	mAP(%)
Holistic + Pose [56]	57.9
Video Darwin [21]	72.0
Interaction Part Mining [85]	72.4
P-CNN [14]	62.3
P-CNN + IDT-FV [14]	71.4
Semantic Features [84]	70.5
Hierarchical Mid-Level Actions [41]	66.8
Higher-order Pooling [13]	73.1
KCP	66.1
BKCP	68.0
B-KCP + KCP	68.6
KCP + Trajectories	73.5
BKCP + Trajectories	72.4
BKCP + KCP + Trajectories	74.7

Table 7 MPII Cooking Activities (7-splits)

ture and *spice*, that involves subtle motions, benefit significantly from SMAID images, while actions such as *fill water from tap* and *open/close drawer* that involve several concurrent sub-actions, benefit most from using KCP. Qualitative SMAID images from the MPII cooking activities and the JHMDB dataset are provided in Figure 7 using a three-channel SMAID, each channel using 7 frames.

In Figures 8 and 9, we provide the accuracy of each class when using KCP as against those from a recent state of the art on the JHMDB and the MPII Cooking activities datasets respectively. On the MPII dataset, we outperform Zhou et al. [85] on 28 sequences (out of 65), and in most cases the improvement is substantial (10-20 sequences). On the JHMDB dataset, we outperform the method in Chéron et al. [14] on 12 sequences against the 21 actions in the dataset.

Algorithm	Avg. Acc. (%)
P-CNN [14]	61.1
P-CNN + IDT-FV [14]	72.2
Action Tubes [23]	62.5
Stacked Fisher Vectors [50]	69.03
IDT + FV [71]	62.8
Higher-order Pooling [13]	73.3
KCP	72.7
BKCP	72.4
BKCP + KCP	73.7
KCP + IDT-FV	74.1
BKCP + KCP + IDT-FV	77.3

Table 8 JHMDB Dataset (3-splits)

Algorithm	HMDB-51(%)	UCF101(%)
Two-stream [60]	59.4	88.0
Very Deep Two-stream Fusion [19]	69.2	93.5
Temporal segment networks[75]	69.4	94.2
Composite LSTM [63]	44.0	84.3
IDT+FV [71]	57.2	85.9
IDT+HFV [51]	61.1	87.9
TDD+IDT [74]	65.9	91.5
DT+MVSF [6]	55.9	83.5
Dynamic Image + IDT-FV [3]	65.2	89.1
Dynamic Flow + IDT-FV[73]	67.4	91.3
Spatial-temporal ResNet [18]	70.3	94.6
KCP	65.8	89.1
BKCP	68.5	88.6
KCP + BKCP	67.8	89.4
KCP + IDT-FV	67.2	92.0
BKCP + IDT-FV	69.6	89.3
BKCP + KCP + IDT-FV	70.5	92.4

Table 9 Average classification accuracy (%) over 3-splits on the HMDB-51 and UCF-101 Datasets.

9 Conclusions

In this paper, we proposed a temporal pooling scheme, *temporal correlation pooling*, based on the correlations between

temporal evolution of classifier scores. Our descriptors are positive definite matrices, thus allowing the use rich mathematical (Riemannian) geometries for non-linear feature pooling. While, our basic descriptor scales quadratically against the number of action classes, we proposed a simple approximation to it, that could scale linearly. We also proposed a novel sub-sequence representation, SMAID, that could increase the temporal receptive field of CNN, thereby improving action classification performance. Using SMAID and temporal correlation pooling schemes, we proposed a novel three-stream end-to-end learnable CNN architecture for action classification. The utility of each of our contributions were substantiated via experiments on four challenging action recognition benchmarks.

Acknowledgements: This research was supported by the Australian Research Council (ARC) through the Centre of Excellence for Robotic Vision (CE140100016) and was undertaken with the resources from the National Computational Infrastructure (NCI) at the Australian National University. The authors also thank Mr. Edison Guo (ANU) for helpful discussions.

References

1. Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 56(2):411–421, 2006.
2. Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *Human Behavior Understanding*, pages 29–39. 2011.
3. Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic image networks for action recognition. In *CVPR*, 2016.
4. Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *ICCV*. IEEE, 2005.
5. Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*. 2014.
6. Zhuowei Cai, Limin Wang, Xiaojiang Peng, and Yu Qiao. Multi-view super vector for action recognition. In *CVPR*, 2014.
7. Joao Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*. 2012.
8. Jose M Chaquet, Enrique J Carmona, and Antonio Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.
9. Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.
10. Xianjie Chen and Alan L Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014.
11. Anoop Cherian, Suvrit Sra, Arindam Banerjee, and Nikolaos Papanikolopoulos. Jensen-bregman logdet divergence with application to efficient similarity search for covariance matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2161–2174, 2013.
12. Anoop Cherian, Basura Fernando, Mehrtash Harandi, and Stephen Gould. Generalized rank pooling for action recognition. In *CVPR*, 2017.
13. Anoop Cherian, Piotr Koniusz, and Stephen Gould. Higher-order pooling of cnn features via kernel linearization for action recognition. In *WACV*, 2017.
14. Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. *arXiv preprint arXiv:1506.03607*, 2015.
15. James W Davis and Aaron F Bobick. The representation and recognition of human movement using temporal templates. In *CVPR*. IEEE, 1997.
16. Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.
17. Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis Bach, and Jean Ponce. Automatic annotation of human actions in video. In *ICCV*, 2009.
18. Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. Spatiotemporal residual networks for video action recognition. In *NIPS*, 2016.
19. Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. 2016.
20. Basura Fernando, Efstratios Gavves, Jose M. Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, 2015.
21. Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, 2015.
22. Juergen Gall, Angela Yao, Negin Razavi, Luc Van Gool, and Victor Lempitsky. Hough forests for object detection, tracking, and action recognition. *PAMI*, 33(11): 2188–2202, 2011.
23. Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 759–768, 2015.

24. Kai Guo, Prakash Ishwar, and Janusz Konrad. Action recognition from video using feature covariance matrices. *TIP*, 2013.
25. Yun He, Soma Shirakabe, Yutaka Satoh, and Hirokatsu Kataoka. Human action recognition without human. *CoRR*, abs/1608.07876, 2016.
26. Samitha Herath, Mehrtaash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and Vision Computing*, 60:4 – 21, 2017. ISSN 0262-8856. Regularization Techniques for High-Dimensional Data Analysis.
27. Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.
28. Zhiwu Huang and Luc Van Gool. A riemannian network for spd matrix learning. In *AAAI*, 2017.
29. Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *ICCV*, 2015.
30. Tony Jebara and Risi Kondor. Bhattacharyya and expected likelihood kernels. In *Learning theory and kernel machines*, pages 57–71. Springer, 2003.
31. H. Jégou, M. Douze, and C. Schmid. On the Burstiness of Visual Elements. *CVPR*, pages 1169–1176, 2009.
32. Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2011.
33. Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, 2013.
34. Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *PAMI*, 35(1):221–231, 2013.
35. Andrej Karpathy, George Toderici, Sachin Shetty, Tommy Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.
36. Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.
37. Piotr Koniusz, Anoop Cherian, and Fatih Porikli. Tensor representations via kernel linearization for action recognition from 3d skeletons. 2016.
38. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
39. Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*. IEEE, 2011.
40. Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *ECCV*. 2014.
41. Tian Lan, Yuke Zhu, Amir Roshan Zamir, and Silvio Savarese. Action recognition by hierarchical mid-level action elements. In *ICCV*, 2015.
42. Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.
43. Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011.
44. Jinna Lei, Xiaofeng Ren, and Dieter Fox. Fine-grained kitchen activity recognition using RGB-D. In *ACM Conference on Ubiquitous Computing*, 2012.
45. Peihua Li, Qilong Wang, Wangmeng Zuo, and Lei Zhang. Log-euclidean kernels for sparse representation and dictionary learning. In *ICCV*, 2013.
46. Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*. Springer, 2016.
47. Bingbing Ni, Vignesh R Paramathayalan, and Philippe Moulin. Multiple granularity analysis for fine-grained action detection. In *CVPR*, 2014.
48. Dan Oneata, Jakob Verbeek, and Cordelia Schmid. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, 2013.
49. Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *ICML*, 2013.
50. Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng. Action recognition with stacked fisher vectors. In *European Conference on Computer Vision*, pages 581–595. Springer, 2014.
51. Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *CVIU*, 2016.
52. Xavier Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.
53. Hamed Pirsiavash and Deva Ramanan. Parsing videos of actions with segmental grammars. In *CVPR*, 2014.
54. Leonid Pishchulin, Mykhaylo Andriluka, and Bernt Schiele. Fine-grained activity recognition with holistic and pose based features. In *Pattern Recognition*, pages 678–689. Springer, 2014.
55. Alessandro Prest, Cordelia Schmid, and Vittorio Ferrari. Weakly supervised learning of interactions between humans and objects. *PAMI*, 34(3):601–614, 2012.

56. Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.
57. Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *arXiv preprint arXiv:1502.06648*, 2015.
58. Michael S Ryoo and Jake K Aggarwal. Recognition of composite human activities through context-free grammar based representation. In *CVPR*, 2006.
59. Sreemananth Sadanand and Jason J Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.
60. Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.
61. Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
62. Suvrit Sra. Positive definite matrices and the symmetric stein divergence. Technical report, 2011.
63. Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, pages 843–852, 2015.
64. Chen Sun and Ram Nevatia. Discover: Discovering important segments for classification of video events and recounting. In *CVPR*, 2014.
65. Kevin Tang, Li Fei-Fei, and Daphne Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.
66. Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christopher Bregler. Efficient object localization using convolutional networks. 2015.
67. Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.
68. Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.
69. Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):480–492, 2012.
70. Chunyu Wang, Yizhou Wang, and Alan L Yuille. An approach to pose-based action recognition. In *CVPR*, 2013.
71. Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.
72. Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1): 60–79, 2013.
73. Jue Wang, Anoop Cherian, and Fatih Porikli. Ordered pooling of optical flow sequences for action recognition. In *WACV*, 2017.
74. Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, 2015.
75. Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.
76. Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.
77. Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *CVPR*, 2015.
78. Angela Yao, Juergen Gall, Gabriele Fanelli, and Luc J Van Gool. Does human action recognition benefit from pose estimation?. In *BMVC*, 2011.
79. Bangpeng Yao and Li Fei-Fei. Action recognition with exemplar based 2.5 d graph matching. In *ECCV*. 2012.
80. Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.
81. Kaicheng Yu and Mathieu Salzmann. Second-order convolutional neural networks. *arXiv preprint arXiv:1703.06817*, 2017.
82. Chunfeng Yuan, Weiming Hu, Xi Li, Stephen Maybank, and Guan Luo. Human action recognition under log-euclidean riemannian metric. In *ACCV*. 2009.
83. Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015.
84. Yang Zhou, Bingbing Ni, Shuicheng Yan, Pierre Moulin, and Qi Tian. Pipelining localized semantic features for fine-grained action recognition. In *ECCV*. 2014.
85. Yang Zhou, Bingbing Ni, Richang Hong, Meng Wang, and Qi Tian. Interaction part mining: A mid-level approach for fine-grained action recognition. In *CVPR*, 2015.
86. Silvia Zuffi and Michael J Black. Puppet flow. *IJCV*, 101(3):437–458, 2013.